Universidade de Brasília
Departamento de Estatística

Um Modelo de Regressão log-Weibull com Fração de Cura para Dados de Pacientes com Aids

por

Letícia Valéria Porfírio

Monografia apresentada para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília
2015

Letícia Valéria Porfírio

Um Modelo de Regressão log-Weibull com Fração de Cura para Dados de Pacientes com Aids

Orientadora:
Prof.ª Dr.ª **Juliana Betini Fachini Gomes**

Monografia apresentada para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília
2015

Dedicatória

À Deus, meu Amado e razão de minha existência,
e à Virgem Maria, minha doce advogada.

À minha querida mãe,
Lucélia Valério Porfírio *por sempre acreditar em mim e dedicar a vida aos filhos, lutando junto a eles em busca dos sonhos de cada um.*

À meu querido pai,
Erasmo de Assis Porfírio*, por ser um exemplo como profissional. Por não medir esforços para que eu concluísse mais esta etapa com todo conforto e êxito.*

Agradecimentos

À Deus, que me conduz no caminho da Verdade.

À professora, orientadora e amiga Juliana Betini Fachini, pela dedicação, paciência e preocupação desde o primeiro momento. Por auxiliar muito mais do que eu esperava.

A todos os professores da UnB que contribuíram para a minha formação.

Aos meus amados pais, por todo apoio e orações diárias.

À minha avó Maria Valério Sobrinho, por toda a preocupação e cuidado comigo.

À minha irmã Hellen Valéria Porfírio, por me entender e estar sempre ao meu lado. Por ser mais que uma irmã, e por se preocupar com minha vida em Cristo.

Aos meus irmãos, Erik Valério Porfírio e Vinícius Valério Porfírio pelo companheirismo e paciência em me aturar nos momentos difíceis.

Ao meu companheiro Marcos Vinícius Freitas Moraes, por sempre enxergar o melhor em mim. Por ser inspiração de força e perseverança.

Aos amigos e colegas de curso que contribuíram para a minha formação, em especial aos "melhores amigos da EST", Brenda Ribeiro, Camila Theresa, João Gustavo, Victor Scatolin, Lena Luciane e Thaís Alvares, pela amizade, pelo apoio, pelas risadas e por estarem comigo desde o início.

À amiga Cecília Évanne de Almeida Miranda, pela ajuda, pela amizade verdadeira desde sempre e pela compreensão nos momentos de ausência.

Ao amigo André Silva de Queiroz, por me ajudar de uma forma que nunca imaginei. Por toda e enorme paciência, e por me mostrar os charmes da estatística.

A todos que de alguma forma contribuíram para a realização deste trabalho.

"Se não puder se destacar pelo talento, vença pelo esforço."
Dave Weinbaum

Sumário

Resumo

Um Modelo de Regressão log-Weibull com Fração de Cura para Dados de Pacientes com Aids

Este trabalho apresenta uma modelagem utilizando técnicas da análise de sobrevivência. Foi estudada a distribuição log-Weibull com fração de cura, para avaliar a evolução clínica de pacientes portadores do vírus HIV em uso da terapia anti-retroviral (HAART), em função da sua adesão ao tratamento. Os parâmetros do modelo foram estimados numericamente pelo método de máxima verossimilhança sujeito a restrição no espaço paramétrico, e ao final foi feita uma análise de resíduos para verificar a adequação do modelo aos dados. Vale ressaltar que, através dos resultados obtidos, surgiu um novo modelo candidato, e assim, utilizou-se o teste da razão de verossimilhanças e as análises gráficas para a comparação entre esses modelos.

Palavras-chave: Análise de Sobrevivência; log-Weibull; Fração de Cura; Verossimilhança; Resíduos.

Abstract

A log-Weibull Regression Model with Healing Fraction for Patients with AIDS Data

This project presents a model using techniques of survival analysis. The log-Weibull distribution with healing fraction was studied to analyze the clinical course of patients with HIV in use of antiretroviral therapy (HAART), according to their adherence to treatment. The parameters of the model were numerically estimated using the method of maximum likelihood subject to parameter space constraints, and at the end of the project, an analysis of residuals was made to verify the adequacy of the model to the data. It is noteworthy that, an alternative model appeared through the results achieved, and thus the likelihood-ratio test and graphical analyzes were used to make a comparison between these two models.

Keywords: Survival analysis; log-Weibull; Healing fraction; likelihood; Residues.

Capítulo 1

Introdução

Em diversas áreas, em especial na área médica, é comum trabalhar com estudos que envolvem as técnicas de análise de sobrevivência, ou seja, estudos longitudinais em que a variável resposta é, geralmente, o tempo até a ocorrência de um evento de interesse. Este tempo é denominado tempo de falha, podendo ser o tempo até a morte do paciente, bem como até a cura ou recidiva de uma doença. As mesmas técnicas de análise de dados são adequadas para aplicações em áreas como, segurança pública, no qual criminalistas estudam o tempo entre a liberação de presos e a ocorrência de crimes; confiabilidade industrial, em que produtos ou componentes são colocados sob teste para se estimar, por exemplo, a probabilidade de certo produto durar mais do que cinco anos; análises financeiras, ciências sociais em geral, entre outras.

Contudo, há estudos em que o evento de interesse não ocorrerá para todos os indivíduos, resultando em observações incompletas, ditas censuradas. Esse tipo de dado pode ser gerado por uma infinidade de circunstâncias, como por exemplo: a saída do paciente do estudo por algum motivo diferente do evento de interesse; a morte do paciente por alguma razão diferente da esperada. A análise de sobrevivência tem como característica principal incorporar a informação proveniente desses dados. Essa característica nos dados é muito importante e é ela que diferencia a análise de sobrevivência das demais técnicas. Então, para esse tipo de estudo, é necessário definir o evento de interesse para que se possam obter as duas informações essenciais para realizar a análise: o tempo de falha ou tempo até a ocorrência do evento de interesse e o tempo de censura.

Sendo assim, foi realizado um estudo com pacientes portadores de AIDS assistidos no do Instituto de Pesquisa Clínica Evandro Chagas (IPEC)/Fiocruz entre janeiro de 2006 e dezembro de 2008 (Campos, 2009), com o intuito de avaliar a evolução clínica dos pacientes em uso da terapia anti-retroviral altamente potente (HAART) em função da sua adesão ao tratamento. A adesão ao tratamento anti-retroviral é considerada fundamental para obter supressão viral e melhora imunológica nos portadores do vírus HIV, e a não

adesão é uma das causas principais de falha terapêutica. Com isso, os dados sobre o fornecimento de medicamentos (dispensas) foram provenientes do Sistema de Controle Logístico de Medicamentos (SICLOM), e o evento de interesse definido nesse estudo foi o tempo decorrido entre a primeira dispensa de medicamentos observada e a ocorrência da falha virológica, que é um dos tipos de falha terapêutica.

Além disso, este estudo traz uma certa particularidade da análise de sobrevivência que é a chamada fração de cura, que basicamente consiste naqueles indivíduos que nunca vão experimentar o evento de interesse, ou seja, indivíduos imunes. Esses indivíduos não suscetíveis ao evento de interesse, aqui, são aqueles em que a adesão ao medicamento se mostra suficientemente eficiente a ponto de impedir a falha virológica. Com isso, pode-se dizer que este trabalho tem como principal objetivo a aplicação das técnicas de análise de sobrevivência, para encontrar um modelo probabilístico com a particularidade de fração de cura, que melhor se ajuste a esses dados.

Vale ressaltar ainda que, assim como em outras áreas, na análise de sobrevivência pode-se estudar o comportamento do tempo quando inclui-se outras covariáveis no estudo. Pois, a variável tempo pode estar sendo influenciada por essas covariáveis da população estudada. Para estudar o efeito dessas covariáveis é necessário então, aplicar um modelo de regressão capaz de acomodar dados censurados. Por isso, os objetivos específicos deste trabalho, são primeiramente, estimar os parâmetros do modelo probabilístico pelo método de máxima verossimilhança, desenvolver um modelo de regressão apropriado e realizar uma análise de resíduos para verificar a adequabilidade deste modelo ajustado.

Todas as análises foram feitas com o auxílio do *software* R (R Core Team, 2015).

Capítulo 2

Revisão de Literatura

2.1 Conceitos básicos

A análise de sobrevivência é uma técnica estatística adequada para analisar dados de acompanhamento de indivíduos ao longo do tempo até a ocorrência de um determinado evento de interesse. Contudo, há estudos em que o evento de interesse não ocorrerá para todos os indivíduos, resultando em observações incompletas, ditas censuradas. Estes dois componentes constituem a resposta. Em estudos clínicos, um conjunto de covariáveis é também, geralmente, medido em cada paciente. Esses e outros conceitos são discutidos em detalhes a seguir.

2.1.1 Tempo de Falha

Segundo Colosimo e Giolo (2006) o tempo de falha é o tempo transcorrido entre a entrada do indivíduo no estudo e a ocorrência do evento de interesse. Esta variável deve estar claramente definida e, portanto, os três elementos que à constituem também, são eles: o tempo inicial. a escala de medida e o evento de interesse (falha).

O tempo inicial do estudo precisa ser bem definido, pois este tempo de origem é utilizado para o cálculo do tempo até a falha. Porém, há estudos em que o indivíduo não participa do estudo a partir da sua data inicial, ou seja, os pacientes entram no estudo em momentos diferentes, este tipo de característica na pesquisa é chamada de coorte aberta, e é também um dos atributos deste trabalho.

A escala de medida é quase sempre o tempo real ou "de relógio". Pode ser medida em dias, horas, semanas, entre outros. É importante definir a escala de medida utilizada para garantir que os tempos até a falha ou censura de todos os indivíduos sejam medidos na mesma escala.

Por fim, é importante, em estudos de sobrevivência, definir de forma clara e precisa o que vem a ser a falha. Se a falha estiver bem definida não haverá dúvidas quanto à natureza do evento que ocorreu com o indivíduo, se foi falha ou se foi censura. Se as definições de falha e censura forem confundidas, a coleta das informações poderá conter erros, comprometendo a qualidade dos resultados da análise de sobrevivência.

O evento de interesse pode ainda ocorrer devido a uma única causa ou devido a duas ou mais. Situações em que causas de falha competem entre si são denominadas na literatura de riscos competitivos, ver Prentice et al. (1978). Esse caso não será abordado neste trabalho.

2.1.2 Censura

Dados censurados são aqueles onde se tem apenas a observação parcial da resposta, isto é, apenas uma observação parcial do tempo de falha. Esse tipo de dado pode ser gerado por uma infinidade de circunstâncias, como por exemplo, a saída do paciente do estudo por algum motivo diferente do evento de interesse, bem como a mudança de residência inviabilizando sua participação no estudo ou a morte do paciente por alguma razão diferente da esperada.

Mesmo censurados, todos os resultados provenientes de um estudo de sobrevivência devem ser utilizados na análise estatística, pois mesmo sendo incompletas, as observações censuradas fornecem informações sobre o tempo de vida de pacientes e, além disso, a omissão das censuras no cálculo das estatísticas de interesse pode acarretar conclusões enviesadas.

Existem alguns tipos de censura, tais como a censura à esquerda, censura intervalar e censura à direita. Neste trabalho a ênfase é dada à censura à direita, que diz respeito à observação parcial da resposta quando o tempo de ocorrência do evento de interesse está à direita do tempo registrado. Além disso, para esse tipo de censura, alguns mecanismos são diferenciados em estudos clínicos. Tem-se que as censuras do tipo I ocorrem naqueles estudos que ao serem finalizados após um período pré-estabelecido de tempo registram, em seu término, alguns indivíduos que não apresentaram o evento de interesse. As censuras do tipo II resultam de estudos os quais são finalizados após a ocorrência do evento de interesse em um número pré-estabelecido de indivíduos. E o terceiro mecanismo de censura, o do tipo aleatório, é o que mais ocorre na prática, é também o tipo de censura que será considerada neste trabalho. Isto acontece quando um paciente é retirado no decorrer do estudo sem ter ocorrido a falha, ou também, por exemplo, se o paciente morrer por uma razão diferente da estudada.

A Figura 2.1 ilustra os mecanismos de censura descritos.

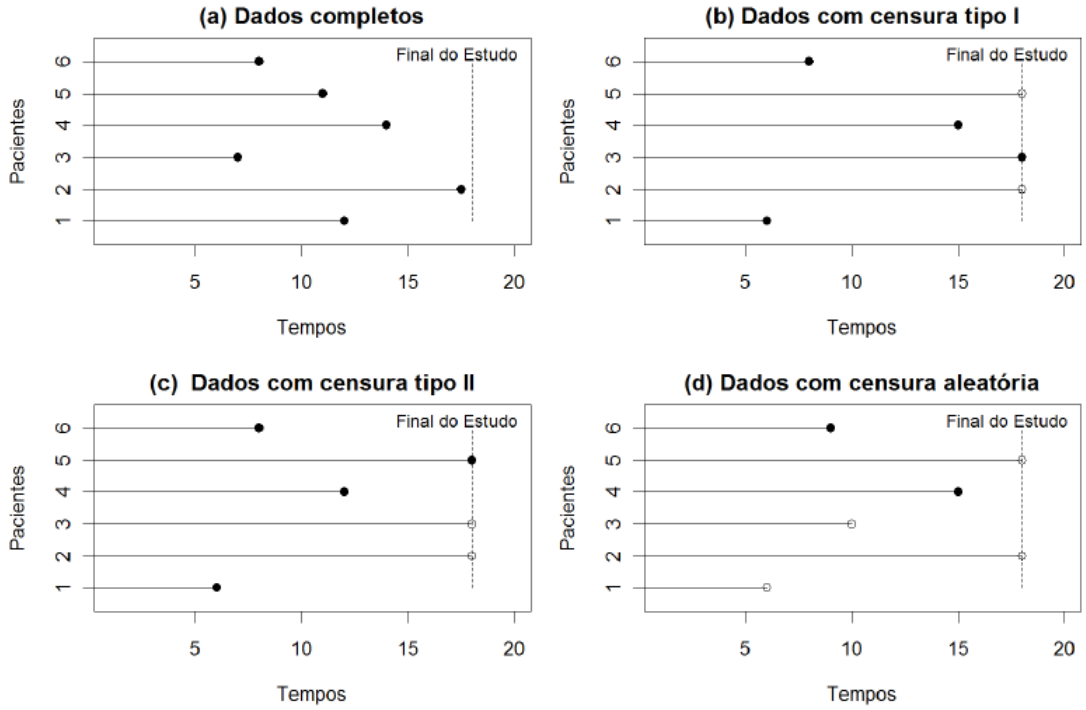

Figura 2.1: Ilustração dos mecanismos de censura em que • representa a falha e ∘ a censura. Fonte: Colosimo e Giolo (2006).

2.2 Representando o Tempo de Sobrevivência

Os principais componentes da análise descritiva envolvendo dados de tempo de vida é a função de sobrevivência $S(t)$ e a função de taxa de falha $\lambda(t)$. Para o cálculo destas funções, define-se o tempo de falha ou censura como a variável aleatória T. Note que T é uma variável não negativa e, em geral, do tipo contínua.

2.2.1 Função de Sobrevivência

A função de sobrevivência é definida como a probabilidade de uma observação não falhar até certo tempo t, ou seja, a probabilidade de um indivíduo sobreviver ao tempo t. Em termos probabilísticos é definida como:

$$S(t) = P(T \geq t).$$

A função de sobrevivência também pode ser encontrada utilizando a função acumulada $F(t)$. Isto é,

$$S(t) = 1 - F(t).$$

Além disso, a função $S(t)$ é uma função monótona, decrescente e contínua (Lawless, 2003).

2.2.2 Função de Taxa de falha ou Função de Risco

A função de taxa de falha ou função de risco é calculada através da razão entre a probabilidade de uma falha ocorrer em um determinado intervalo de tempo $[t, t + \Delta t)$, dado que não tenha acontecido falha até o instante de tempo t, e o tamanho do intervalo. Ou seja,

$$\lambda(t) = \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}.$$

Assumindo Δt bem pequeno, o intervalo de tempo será tão pequeno que a função de taxa de falha pode ser considerada a taxa de falha instantânea para um indivíduo que sobreviveu ao tempo t. A taxa de falha é útil para descrever a distribuição do tempo de vida de pacientes. Ela representa a forma em que a taxa instantânea de falha, ou força de mortalidade, muda com o tempo.

A função de risco é mais informativa do que a função de sobrevivência. Diferentes funções de sobrevivência podem ter formas semelhantes, enquanto as respectivas funções de taxa de falha podem diferir drasticamente. Dessa forma, a modelagem da função taxa de falha é um importante método para dados de sobrevivência, pois pode ter forma crescente, decrescente, ou constante, unimodal ou em forma de banheira, como mostra a Figura 2.2.

A função crescente indica que a taxa de falha do paciente aumenta com o transcorrer do tempo, mostrando que existe um efeito gradual de envelhecimento. A função constante indica que a taxa de falha não se altera com o passar do tempo. A função decrescente mostra que a taxa de falha diminui à medida que o tempo passa.

Outra função que pode ser usada em análise de sobrevivência é a função taxa de falha acumulada, que é definida por:

$$\Lambda(t) = \int_0^t \lambda(u) du.$$

Essa função não tem interpretação direta, mas é útil na avaliação da função de maior interesse que é taxa de falha, $\lambda(t)$. Isto acontece na estimação não-paramétrica em que $\Lambda(t)$ apresenta um estimador com propriedades ótimas e $\lambda(t)$ é difícil de ser estimada (Colosimo e Giolo, 2006).

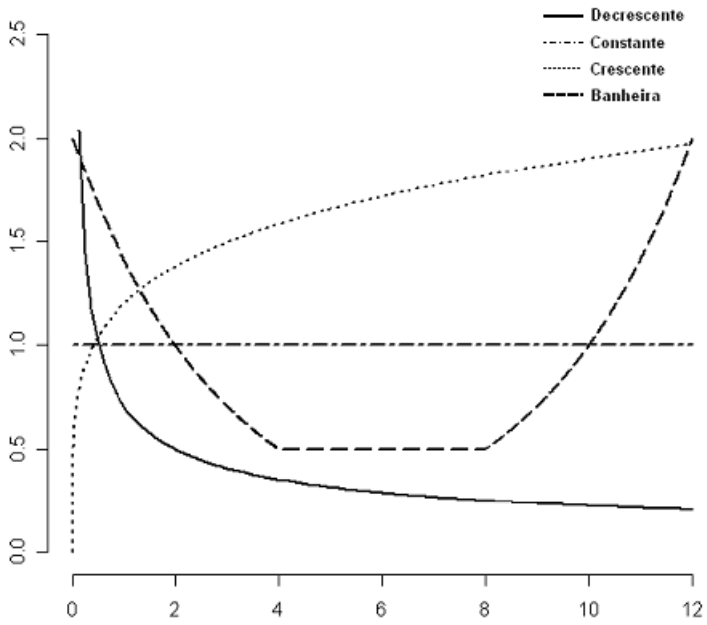

Figura 2.2: Funções de Taxa de Falha. Fonte: portalaction.com.br (2015).

2.2.3 Relações entre as Funções

Uma interessante característica entre a função de sobrevivência, a função de taxa de falha, a função de taxa de falha acumulada e a função densidade de probabilidade, é que elas são matematicamente relacionadas. Essa relação pode ser útil nos processos de estimação ou em situações que se conhece uma das funções e deseja-se obter a outra função. Por exemplo, há situações em que a estimativa de uma função será mais fácil de ser obtida do que a estimativa de outra função, então, neste caso, estima-se a função que for mais simples de ser calculada e depois encontra-se a outra função através da relação entre elas.

Essas relações são expressas por:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\Big(\log S(t)\Big),$$

$$\Lambda(t) = \int_0^t \lambda(u)du = -\log S(t)$$

e

$$S(t) = \exp\{-\Lambda(t)\} = \exp\left\{-\int_0^t \lambda(u)du\right\}.$$

2.3 Técnicas Não-Paramétricas

Nos textos básicos de estatística, uma análise descritiva consiste essencialmente em encontrar medidas de tendência central e variabilidade. Como a presença de censuras invalida este tipo de tratamento aos dados de sobrevivência, as principais componentes da análise descritiva envolvendo dados de tempo de vida são a função de sobrevivência $S(t)$, e a função de taxa de falha $\lambda(t)$ abordadas anteriormente.

Esta seção inicialmente apresenta o conhecido estimador não-paramétrico de Kaplan-Meier para a função de sobrevivência, que também pode ser estimada por outros estimadores como o de Nelson-Aalen e a tabela de vida, porém, estes não serão abordados neste trabalho devido a superioridade do estimador de Kaplan-Meier.

Por fim, é abordado aqui um outro método não-paramétrico que estuda formas da função de taxa de falha para ajudar na modelagem do estudo.

2.3.1 O Estimador de Kaplan-Meier

O estimador Kaplan-Meier, também conhecido por estimador limite-produto, é um estimador não-paramétrico da função de sobrevivência. Este é sem dúvida o mais utilizado em estudos clínicos e vem ganhando cada vez mais espaço em estudos de confiabilidade. Ele é uma adaptação da função de sobrevivência empírica que, na ausência de censuras, é definida como:

$$\widehat{S}(t) = \frac{\text{nº de observações que não falharam até o tempo } t}{\text{nº de observações no estudo}}. \tag{2.1}$$

Será definida a seguir a expressão geral deste estimador, assim como foi proposto por seus autores.

A expressão geral do Kaplan-Meier pode ser apresentada após estas considerações preliminares:

- $t_1 < t_2 < \ldots < t_k$, os k tempos distintos e ordenados de falha;
- d_j o número falhas em $t_j, j = 1, 2, \ldots, k$;
- n_j o número de indivíduos sob risco em t_j, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j.

Assim, o estimador Kaplan-Meier é definido como:

$$\widehat{S}(t) = \prod_{j\,:\,t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j\,:\,t_j<t} \left(1 - \frac{d_j}{n_j} \right). \tag{2.2}$$

As principais propriedades do estimador de Kaplan-Meier são basicamente as seguintes:

- é não-viciado para amostras grandes,
- é fracamente consistente,
- converge assintoticamente para um processo gaussiano e
- é estimador de máxima verossimilhança de $S(t)$.

Este estimador tem como característica ser uma função escada, onde os "degraus" ocorrem nos instantes de tempo t em que ocorrem as falhas. Serão tantos intervalos quanto forem o número de falhas distintas.

Segundo Colosimo e Giolo (2006), para que se possa construir intervalos de confiança e testar hipóteses para $S(t)$, é necessário, no entanto, avaliar a precisão de estimador. A expressão para a variância assintótica do estimador de Kaplan-Meier é dada por:

$$\widehat{Var}\big(\widehat{S}(t)\big) = \Big[\widehat{S}(t)\Big]^2 \sum_{j\,:\,t_j<t} \frac{d_j}{n_j(n_j - d_j)}. \tag{2.3}$$

Um intervalo de confiança para $S(t)$ considerando $100(1-\alpha)\%$ de confiança é dado por:

$$\widehat{S}(t) \pm z_{\alpha/2}\sqrt{\widehat{Var}\big(\widehat{S}(t)\big)}, \tag{2.4}$$

onde $\alpha/2$ denota o $\alpha/2$-percentil da distribuição Normal padrão.

Entretanto, deve-se observar que o intervalo de confiança de $S(t)$, dado pela equação 2.4, não é muito indicado para amostras pequenas e, se t é um valor extremo, pode incluir valores fora do intervalo (0,1). Uma alternativa é utilizar uma transformação para $S(t)$, como por exemplo, $\widehat{U}(t) = \log[-\log(\widehat{S}(t))]$, sugerida por Kalbfleisch e Prentice (2002).

2.3.2 Determinação Empírica da Forma da Função de Risco

Como existem várias formas que o gráfico da função de taxa de falha da variável T pode assumir, é importante utilizar uma metodologia para identificar o modelo mais

apropriado para esta variável. Uma técnica de verificação gráfica para ajuste do modelo é conhecida como Curva do Tempo Total em Teste (TTT plot), proposto por Aarset (1987).

A curva TTT é obtida construindo um gráfico de

$$G(r/n) = \frac{[(\sum_{i=1}^{r} T_{1:n}) + (n-r)T_{1:n})]}{(\sum_{i=1}^{n} T_{1:n})},$$

por r/n, sendo que $r = 1, ..., n$, e $T_{i:n}, i = 1, ..., n$ são as estatísticas de ordem da amostra.

A Figura 2.3 ilustra as diferentes formas que a curva TTT plot pode assumir. Sendo essas:

- Reta diagonal (**A**) ⇒ Função taxa de falha constante é adequada.
- Curva convexa (**B**) ou côncava (**C**) ⇒ Função taxa de falha é monotonicamente decrescente ou crescente, respectivamente.
- Curva convexa e depois côncava (**D**) ⇒ Função taxa de falha tem forma de **U**.
- Curva côncava e depois convexa (**E**) ⇒ Função taxa de falha é unimodal.

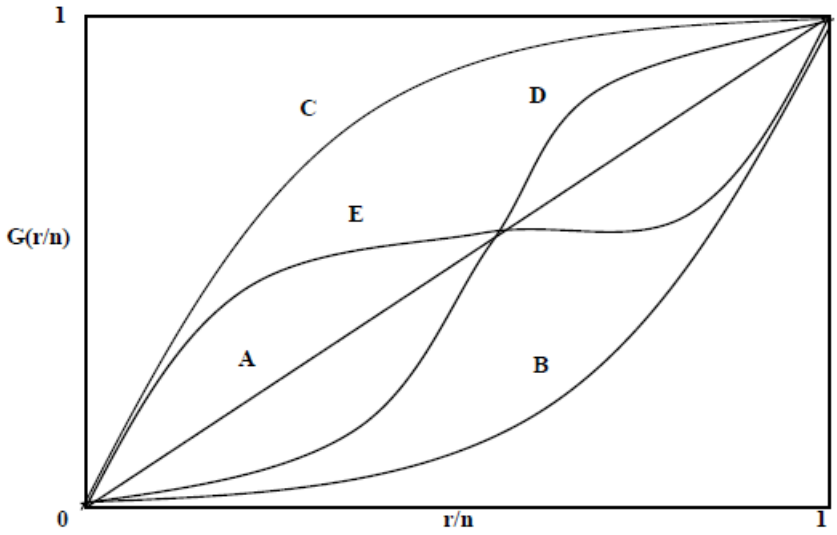

Figura 2.3: Diferentes formas da Curva TTT plot.

2.4 Modelos Probabilísticos

Sendo T a variável aleatória que define o tempo de falha ou censura dos dados de sobrevivência, o próximo passo a se pensar seria encontrar uma distribuição de probabilidade que melhor a represente. Existem na literatura várias distribuições de probabilidade

que têm particularidades diferentes, o que torna cada distribuição mais adequada a um tipo diferente de variável aleatória.

A primeira característica de T que orienta a escolha de uma distribuição de probabilidade é que esta é uma quantidade contínua e não negativa, devido a essa particularidade, as distribuições de probabilidade que conseguem modelar os dados de sobrevivência são distribuições em que a variável aleatória é definida para valores maiores ou iguais a zero. Além disso, frequentemente o tempo de sobrevivência T apresenta forte assimetria, com uma grande cauda à direita. Este formato assimétrico decorre de observarmos grande parte dos tempos de sobrevivência com valores pequenos e poucos com tempos muito longos.

Uma possível distribuição de probabilidade que poderá ser utilizada para modelar os dados desse trabalho é a distribuição Weibull, e essa será definida na Seção 2.4.1.

2.4.1 Distribuição Weibull

A distrição Weibull é atualmente a mais utilizada para modelar tempos de sobrevivência, principalmente na área biomédica, isso se deve ao fato dela apresentar uma grande variedade de formas, todas com uma propriedade básica: a sua função de taxa de falha é monótona, isto é, ela é crescente, decrescente ou constante.

Uma variável aleatória T que segue esta distribuição tem a seguinte função densidade de probabilidade:

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp\left\{ -\left(\frac{t}{\alpha}\right)^\gamma \right\}, \quad t \geq 0, \tag{2.5}$$

em que γ, o parâmetro de forma, e α, o parâmetro de escala, são ambos positivos. O parâmetro α tem a mesma unidade de medida de t e γ não tem unidade de medida.

Para esta distribuição, as funções de sobrevivência e de taxa de falha são, respectivamente,

$$S(t) = \exp\left\{ -\left(\frac{t}{\alpha}\right)^\gamma \right\} \tag{2.6}$$

e

$$\lambda(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1}, \tag{2.7}$$

para $t \geq 0$, α e $\gamma > 0$.

O parâmetro γ determina a forma da função de risco da seguinte maneira:

- $\gamma < 1 \longrightarrow$ a função de risco decresce,
- $\gamma > 1 \longrightarrow$ a função de risco crescente,
- $\gamma = 1 \longrightarrow$ a função de risco é constante. Neste caso, tem-se o caso particular da distribuição Weibull, a distribuição Exponencial.

As expressões da média e variância da distribuição Weibull são mostradas abaixo, elas incluem o uso da função gama, isto é,

$$E(T) = \alpha\Gamma\left[1 + \left(\frac{1}{\gamma}\right)\right],$$

$$Var(T) = \sigma^2\left[\Gamma\left[1 + \left(\frac{2}{\gamma}\right)\right] - \Gamma\left[1 + \left(\frac{1}{\gamma}\right)\right]^2\right],$$

sendo a função gama $\Gamma(\beta)$, definida por:

$$\Gamma(\beta) = \int_0^\infty x^{\beta-1}\exp\{-x\}dx.$$

Na análise de tempos de vida, há situações em que é conveniente utilizar o logaritmo da variável T. Quando esta variável segue distribuição Weibull, o seu logaritmo segue uma distribuição do Valor Extremo ou de Gambel, ou ainda chamada de log-Weibull, que neste trabalho foi a nomenclatura utilizada para a variável Y. Ou seja, se a variável T tem distribuição de Weibull com $f(t)$ dada pela equação (2.5), então, a variável $Y = \log(T)$ tem uma distribuição log-Weibull (ou do valor extremo) com a seguinte função de densidade:

$$f(y) = \frac{1}{\sigma}\exp\left\{\left(\frac{y-\mu}{\sigma}\right) - \exp\left(\frac{y-\mu}{\sigma}\right)\right\}, \tag{2.8}$$

em que y e $\mu \in \Re$ e $\sigma > 0$. A prova está descrita no anexo A.1. Se $\mu = 0$ e $\sigma = 1$ tem-se a distribuição do Valor Extremo Padrão. Os parâmetros μ e σ são denominados parâmetros de locação e escala, respectivamente. Os parâmetros das distribuições Weibull e do Valor Extremo apresentam as seguintes relações de igualdade: $\gamma = \frac{1}{\sigma}$ e $\alpha = \exp\{\mu\}$.

As funções de sobrevivência e de taxa de falha da variável Y são dadas, respectivamente, por:

$$S(y) = \exp\left\{-\exp\left\{\frac{y-\mu}{\sigma}\right\}\right\} \tag{2.9}$$

e

$$\lambda(y) = \frac{1}{\sigma} \exp\left\{\frac{y-\mu}{\sigma}\right\}. \tag{2.10}$$

2.4.2 Estimação dos Parâmetros do Modelo

Um modelo probabilístico, é caracterizado por quantidades desconhecidas, denominadas parâmetros. Essas quantidades conferem uma forma geral aos modelos probabilísticos. Entretanto, em cada estudo envolvendo tempos de falha, os parâmetros devem ser estimados a partir das observações amostrais, para que o modelo fique determinado e, assim, seja possível responder às perguntas de interesse. Os dados observados para uma amostra servem para que estes parâmetros possam ser estimados. (Colosimo e Giolo, 2006).

Dentro do universo estatístico é possível citar alguns importantes métodos de estimação, talvez o mais conhecido seja o método dos mínimos quadrados, no entanto, este método é inapropriado para estudos de tempos de vida, pois não é capaz de incorporar a informação das observações censuradas no seu processo de estimação.

O método de máxima verossimilhança surge como uma opção apropriada para este tipo de dados, uma vez que ele incorpora as censuras, é relativamente simples e possui propriedades ótimas para grandes amostras. Este método é melhor apresentado a seguir.

Método de Máxima Verossimilhança

A estimação dos parâmetros no método de máxima verossimilhança é feita baseada nos resultados obtidos pela amostra, e a ideia desse método é achar a distribuição que tenha a maior probabilidade de ter gerado aquela amostra. Com esse intuito, o método procura os valores de $\boldsymbol{\theta}$ que maximizem a função de verossimilhança $L(\boldsymbol{\theta})$, pois são esses valores que têm a maior probabilidade de ter gerado a amostra observada. A função de verossimilhança para um parâmetro $\boldsymbol{\theta}$ de uma dada população com f.d.p $f(\cdot;\boldsymbol{\theta})$ é, então, expressa por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i;\boldsymbol{\theta}), \tag{2.11}$$

onde n é o número total de observações e $\boldsymbol{\theta}$ o vetor de parâmetros.

Segundo Colosimo e Giolo (2006), a função de verossimilhança $L(\boldsymbol{\theta})$ mostra que a contribuição de cada observação não-censurada é sua função de densidade. A contri-

buição de cada observação censurada não é, contudo, a sua função de densidade. Estas observações somente nos informam que o tempo de falha é maior que o tempo de censura observado e, portanto, que a sua contribuição para $L(\boldsymbol{\theta})$ é a sua função de sobrevivência $S(t)$.

Então, com base no número de observações não-censuradas (r) e nas censuradas $(n-r)$, a função de verossimilhança é apresentada com base nos tipo de censura em estudo, que neste estudo trata-se da censura à direita do tipo aleatória, contudo, como mostra Colosimo e Giolo (2006), a expressão para a função de verossimilhança para todos os mecanismos de censura à direita, a menos de constantes, é a mesma e é dada por:

$$\begin{aligned} L(\boldsymbol{\theta}) &\propto \prod_{i=1}^{n} \Big[f(t_i;\boldsymbol{\theta})\Big]^{\delta_i} \Big[S(t_i;\boldsymbol{\theta})\Big]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \Big[\lambda(t_i;\boldsymbol{\theta})\Big]^{\delta_i} S(t_i;\boldsymbol{\theta}), \end{aligned} \tag{2.12}$$

em que

$$\delta_i = \begin{cases} 0 \text{ se } t_i \text{ é um tempo de falha;} \\ 1 \text{ se } t_i \text{ é um tempo de censura.} \end{cases}$$

É sempre conveniente, no entanto, trabalhar com o logaritmo da função de verossimilhança (2.12). Os estimadores de máxima verossimilhança são os valores de $\boldsymbol{\theta}$ que maximizam $L(\boldsymbol{\theta})$ ou equivalentemente o logaritmo de $L(\boldsymbol{\theta})$, isto é, $\log(L(\boldsymbol{\theta})$. Eles são encontrados resolvendo-se o sistema de equações:

$$U(\boldsymbol{\theta}) = \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0.$$

2.4.3 Intervalo de Confiança

Segundo Colosimo e Giolo (2006), após estimar os parâmetros da distribuição pelo método de máxima verossimilhança, pode-se então calcular o intervalo de confiança dessas estimativas. Uma propriedade importante para a construção de intervalos de confiança é a que diz respeito à distribuição assintótica dos estimadores de máxima verossimilhanças $\widehat{\boldsymbol{\theta}}$. Para grandes amostras, esta propriedade estabelece, sob certas condições de regularidade, que a distribuição do vetor $\widehat{\boldsymbol{\theta}}$ é Normal multivariada de média $\boldsymbol{\theta}$ e matriz de variância-covariância $Var(\widehat{\boldsymbol{\theta}})$, isto é,

$$\widehat{\boldsymbol{\theta}} \sim N_k(\boldsymbol{\theta}, Var(\widehat{\boldsymbol{\theta}})),$$

em que

$$Var(\widehat{\boldsymbol{\theta}}) \approx -[I_F(\boldsymbol{\theta})]^{-1}$$

e

$$I_F = E\left[\left(\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2\right].$$

Devido à dificuldade de calcular a esperança, usa-se a matriz de informação observada avaliada em $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. Portanto, um intervalo de confiança aproximado de $(1-\alpha)100\%$ de confiança para $\boldsymbol{\theta}$ é dado por:

$$\widehat{\boldsymbol{\theta}} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\widehat{\boldsymbol{\theta}})}.$$

Após obter as estimativas e intervalos de confiança, também é interessante testar hipóteses para o vetor de parâmetros $\boldsymbol{\theta}$. Uma das hipóteses que podem ser testadas são:

$$\begin{cases} H_0 : \theta_j = 0; \\ H_1 : \theta_j \neq 0. \end{cases}$$

A distribuição assintótica dos estimadores é a generalização da t de *student*. Para grandes amostras, a t se aproxima da distribuição Normal. Portanto, o teste que é baseado na distribuição assintótica dos estimadores sob a hipótese nula levantada acima tem como estatística do teste:

$$Z = \frac{\widehat{\theta}_j - 0}{\sqrt{\widehat{Var}(\widehat{\theta}_j)}} \sim N(0,1).$$

2.4.4 Teste da Razão de Verossimilhanças

Para um modelo com um vetor $\boldsymbol{\theta}$ de parâmetros, muitas vezes há o interesse em testar hipóteses relacionadas a este vetor ou a um subconjunto dele (Colosimo e Giolo, 2006).

Para o teste da Razão de Verossimilhanças é feita a comparação dos valores dos logaritmos da função de verossimilhança maximizada e sob H_0, ou seja, a comparação de $\log L(\widehat{\boldsymbol{\theta}})$ e $\log L(\widehat{\boldsymbol{\theta}}_0)$. A estatística do teste é dada por:

$$TRV = -2\log\left[\frac{L(\widehat{\boldsymbol{\theta}}_0)}{L(\widehat{\boldsymbol{\theta}})}\right] = 2[\log(L(\widehat{\boldsymbol{\theta}}) - \log L(\widehat{\boldsymbol{\theta}}_0)],$$

em que, sob $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, segue aproximadamente uma distribuição χ^2 com p graus de liberdade, em que p corresponde à diferença entre o número de parâmetros do modelo completo e o número de parâmetros do modelo restrito. Para amostras grandes, H_0 é rejeitada, a um nível $100\alpha\%$ de significância, se $TRV > \chi^2_{p,1-\alpha}$.

Pode-se ainda encontrar na literatura outros testes para os parâmetros do modelos. Os mais comuns são os Teste de Wald e o Teste Escore, descritos brevemente por Colosimo e Giolo (2006).

2.4.5 Fração de Cura

Os principais modelos de análise de sobrevivência têm como pressuposto o fato de que os indivíduos do estudo irão, em algum momento, experimentar o evento de interesse definido.

A função de sobrevivência é considerada própria quando todos os indivíduos são suscetíveis ao evento de interesse. Entretanto, existem situações que, para uma proporção de indivíduos, o evento de interesse não ocorrerá. A função de sobrevivência é considerada imprópria quando à medida que o tempo tende a infinito essa função não tende a zero. Isso pode indicar que existe uma proporção de indivíduos curados ou imunes. Dessa forma, a função de sobrevivência juntamente com suas propriedades são muito importantes para a identificação de dados com a presença de indivíduos curados.

Nesse caso, considerar os modelos de sobrevivência usuais, que assumem que a função de sobrevivência converge para zero quando a variável tempo tende a infinito (função de sobrevivência própria), podem não ser adequados. Para modelar esse tipo de dados, modelos com fração de cura são mais apropriados (Fachini, 2011).

Esses indivíduos não suscetíveis, definidos como curados, aparecem na base de dados como observações censuradas, visto que o evento de interesse não é observado, ou seja, não houve falha.

O indicativo de indivíduos curados na base de dados é verificada ao construir o gráfico da função de sobrevivência empírica, estimada pelo método de Kaplan-Meier. Verifica-se no gráfico o comportamento da cauda direita e caso ela permaneça de maneira constante em um nível acima de zero por um período grande, conclui-se que há indicativo de indivíduos curados (Fachini, 2011). A modelagem de fração de cura para o caso univariado pode ser abordada seguindo a metodologia introduzida por Berkson e Gage (1952) que considera a construção de uma função de sobrevivência populacional na forma de mistura.

Quando a metodologia de proporção de curados introduzida por Berkson e Gage (1952) é adotada para modelar, utiliza-se a função de sobrevivência populacional em forma de mistura, uma vez que a população é dividida em duas partes: indivíduos suscetíveis, e indivíduos não suscetíveis ao evento de interesse. Assim, a função de sobrevivência

populacional em forma de mistura é dada por:

$$S_{pop}(t) = (1 - \phi) + \phi S(t), \tag{2.13}$$

sendo $S(t)$ uma função de sobrevivência própria, podendo ser a função de sobrevivência do modelo Weibull ou do Valor Extremo, ou ainda a função de sobrevivência de outra distribuição de probabilidade, $(1-\phi)$ a probabilidade de indivíduos serem não suscetíveis (curados), e ϕ a probabilidade de indivíduos serem suscetíveis ao evento de interesse.

As propriedades da função são:

- $lim_{t\to\infty} S_{pop}(t) = (1 - \phi)$ e
- $lim_{t\to 0} S_{pop}(t) = 1$.

Usando as relações da função densidade probabilidade, função de sobrevivência e função de risco, a função densidade populacional e a função de risco populacional são respectivamente representadas por:

$$f_{pop}(t) = \phi f(t) \tag{2.14}$$

e

$$h_{pop}(t) = \frac{\phi f(t)}{(1 - \phi) + \phi S(t)}. \tag{2.15}$$

Pode-se notar ainda que a probabilidade de cura $(1-\phi)$ pode variar de indivíduo para indivíduo, pois é razoável assumir que esta pode depender de característica individuais (covariáveis). Nesses casos, o efeito de covariáveis na probabilidade de cura é comumente formulado por um modelo logístico. Segue que a probabilidade de cura pode ser modelada em termos do vetor de covariáveis $\mathbf{x}^T$ por meio da função logística, de modo que:

$$1 - \phi = 1 - \phi(\mathbf{x}^T) = \frac{\exp\left(\mathbf{x}^T\boldsymbol{\lambda}\right)}{1 + \exp\left(\mathbf{x}^T\boldsymbol{\lambda}\right)}, \tag{2.16}$$

porém, as covariáveis também podem ser introduzidas no parâmetro de cura por meio da seguinte relação:

$$1 - \phi = 1 - \phi(\mathbf{x}^T) = \exp\big\{ - \left(\mathbf{x}^T\boldsymbol{\lambda}\right)\big\}, \tag{2.17}$$

em que $\boldsymbol{\lambda}$ denota o vetor de parâmetros de regressão.

A literatura sobre modelos de sobrevivência com fração de cura é extensa e está em rápido desenvolvimento. É possível destacar como referências fundamentais os livros de

Malller e Zhou (1996) e Ibrahim et al. (2001), como também o artigo de Tsodicov et al. (2003) e o artigo de Cooner et al. (2007).

2.5 Modelo de Regressão Locação Escala

Os estudos em análise de sobrevivência muitas vezes envolvem covariáveis que podem estar relacionadas com o tempo de sobrevivência, e certamente essas covariáveis devem ser incluídas na análise estatística dos dados. Neste contexto, a forma mais eficiente de acomodar o efeito dessas covariáveis é utilizar um modelo de regressão apropriado para dados censurados.

Na análise de sobrevivência, existem duas classes importantes de modelos de regressão: os modelos de riscos proporcionais e os modelos de locação e escala (Carrasco, 2007). Neste trabalho é utilizado o modelo de locação escala, em que as covariáveis são inseridas no parâmetro de locação, apesar de também ser possível inserir essas covariáveis diretamente no modelo probabilístico.

Seja $\mathbf{x}^T = (x_1, x_2, ..., x_p)$ um vetor formado por observações de p variáveis regressoras. O modelo, caracterizado por $Y = \log(T)$, ou seja, Y como logaritmo dos tempos de sobrevivência, é caracterizado por:

$$Y = \mu(x) + \sigma z,$$

em que $\sigma > 0$ representa o parâmetro de escala, $-\infty < \mu < +\infty$ o parâmetro de locação, dependendo das variáveis regressoras, e $\mathbf{z}$ é o erro aleatório. Geralmente considera-se $\mu(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}$ em que $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ representa o vetor de parâmetros desconhecidos associado às covariáveis.

O modelo de locação e escala é portanto um modelo log linear para a variável T e as variáveis regressoras atuam multiplicativamente sobre T, por isso esses modelos também podem ser chamados de modelos de tempo de vida acelerado, ou seja, segundo Colosimo e Giolo (2006), o efeito das covariáveis é de acelerar ou desacelerar o tempo de sobrevivência.

2.5.1 Modelo de Regressão Log-Weibull

Aplicando o resultado exposto acima para $Y = \log(T)$ em que T tem distribuição Weibull, ou seja, Y tem distribuição do valor extremo ou de Gambel (ou como denominada aqui, distribuição log-Weibull), como visto na seção 2.4.1, as funções densidade de

probabilidade e de sobrevivência podem ser expressas, respectivamente por:

$$f(y) = \frac{1}{\sigma} \exp\left\{ \left(\frac{y - \mathbf{x}^T\boldsymbol{\beta}}{\sigma} \right) - \exp\left(\frac{y - \mathbf{x}^T\boldsymbol{\beta}}{\sigma} \right) \right\} \tag{2.18}$$

e

$$S(y) = \exp\left\{ - \exp\left\{ \frac{y - \mathbf{x}^T\boldsymbol{\beta}}{\sigma} \right\} \right\}, \tag{2.19}$$

sendo y e $\boldsymbol{\beta} \in \Re$ e $\sigma > 0$.

2.5.2 Adequação do Modelo Ajustado

É parte fundamental da análise dos dados que a adequabilidade do modelo de regressão proposto seja verificada. Análises gráficas de resíduos são geralmente utilizadas para este fim, sendo o maior potencial desse tipo de análise apontar falhas nas suposições, embora elas também sejam úteis na detecção de observações atípicas.

Colosimo e Giolo (2006) propõem a análise de quatro resíduos. A seguir, estão descritos três deles: resíduos de Cox-Snell, que é utilizado para analisar o ajuste global do modelo; resíduos *martingal*, utilizado para determinar a forma funcional de uma covariável, em geral contínua, sendo incluída no modelo de regressão; e os resíduos *deviance*, útil quando se examina a acurácia do modelo para cada indivíduo sob estudo.

Resíduos de Cox-Snell

Esses resíduos são quantidades determinadas por:

$$\widehat{e_i} = \widehat{\Lambda}(t_i|x_i), \tag{2.20}$$

em que $\widehat{\Lambda}(\cdot)$ é a função de taxa de falha acumulada obtida do modelo ajustado.

Segundo Lawless (2003) esse resíduo vem de uma população homogênea e deve seguir uma distribuição exponencial padrão se o modelo for adequado. Desse modo, como dito por Colosimo e Giolo (2006), pode-se fazer uso de técnicas gráficas, ou seja, o gráfico $\widehat{e_i} \times \widehat{\Lambda}(\widehat{e_i})$ deve ser aproximadamente uma reta com inclinação 1, quando o modelo exponencial for adequado, uma vez que $\widehat{\Lambda}(\widehat{e_i}) = -\log(\widehat{S}(\widehat{e_i}))$. Aqui $\widehat{S}(\widehat{e_i})$ é a função de sobrevivência dos $\widehat{e_i}$'s obtida pelo estimador de Kaplan-Meier. O gráfico da curva de sobrevivência desses resíduos, obtidas por Kaplan-Meier e pelo modelo exponencial padrão, também auxiliam na verificação da qualidade do modelo ajustado. Quanto mais

próximas elas se apresentarem, melhor é considerado o ajuste do modelo aos dados.

Resíduos Martingal

Esses resíduos são definidos por:

$$\widehat{m}_i = \delta_i - \widehat{e}_i, \tag{2.21}$$

onde δ_i é a varável indicadora de censura e $\widehat{e}_i$ são os resíduos de Cox-Snell. Como já dito anteriormente, os resíduos *martingal* servem para verificar a melhor forma funcional de uma covariável quantitativa dos dados, por exemplo, se o diagrama de dispersão da covariável x_1 com os resíduos *martingal* apresentar uma relação linear, não será necessário fazer transformação na covariável x_1. Contudo, se o diagrama de dispersão apresentar uma forma de uma parábola, então há indícios de que é necessário acrescentar um termo quadrático da variável x_1 no modelo de regressão.

Resíduos Deviance

Os resíduos *deviance* são definidos por:

$$\widehat{d}_i = \text{sinal}(\widehat{m}_i)\Big[-2\Big(\widehat{m}_i + \delta_i \log(\delta_i - \widehat{m}_i)\Big)\Big]^{\frac{1}{2}}, \tag{2.22}$$

em que, $\widehat{m}_i$ é o resíduo *martingal*.

Esses resíduos, que são uma tentativa de tornar os resíduos *martingal* mais simétricos em torno de zero, facilitam, em geral, a detecção de pontos atípicos (*outliers*). Se o modelo for apropriado, esses resíduos devem apresentar um comportamento aleatório em torno de zero. Gráficos dos resíduos *martingal*, ou *deviance*, versus os tempos fornecem, assim, uma forma de verificar a adequação do modelo ajustado, bem como auxiliam na detecção de observações atípicas (Colosimo e Giolo, 2006).

Capítulo 3

Metodologia

3.1 Material

O conjunto de dados que foi utilizado nesse trabalho é referente ao estudo sobre o efeito da adesão ao tratamento antirretroviral na ocorrência da falha terapêutica (Campos, 2009), mas que aqui neste trabalho, é considerada apenas a falha virológica(carga viral detectável por 6 meses após o início da terapia anti-retroviral altamente potente (HAART) ou carga viral detectável após alcançar a supressão viral), que foi realizado com pacientes portadores do vírus HIV assistidos no Ipec/Fiocruz entre janeiro de 2006 e dezembro de 2008.

A variável de interesse foi obtida a partir do Sistema de Controle Logístico de Medicamentos (SICLOM), desenvolvido para a dispensa de medicamentos antirretrovirais, e calculada como a razão entre o total de dias com atraso no contato com a farmácia para obter a medicação e os dias de acompanhamento entre a entrada no estudo e a falha terapêutica. Como se trata de dados prevalentes utilizou-se a data da primeira dispensa como início do acompanhamento e a data da falha ou da censura como data fim. As variáveis registradas para cada paciente que serão utilizadas neste estudo estão listadas na Tabela 3.1. A base de dados é composta por 711 observações.

Para este estudo foi incluída apenas uma covariável, que é relacionada ao tratamento, sendo esta a quantidade média de comprimidos prescritos por dia para o paciente no período avaliado (os medicamentos injetáveis e as soluções foram considerados como uma dose/comprimido).

Tabela 3.1: Variáveis do banco de dados *adesao.dat.*

Variável	Descrição
id	identificação do paciente
ini	data do início do acompanhamento da dispensação de medicamentos antirretrovirais (em dias)
fim	data da falha virológica ou fim do estudo
tempo	*fim* - *ini* (em dias)
status	0 = censura, 1 = falha virológica
comprimdia	número médio de comprimidos prescritos para o paciente/dia

3.2 Métodos

Modelo de Regressão log-Weibull com Fração de Cura

Nas seções 2.4.5 e 2.5 foi realizado uma revisão de literatura do modelo de regressão log-Weibull e dos modelos de fração de cura. Agora, será descrito o modelo de regressão log-Weibull com fração de cura seguindo a abordagem de Berkson e Gage (1952), objetivo principal deste trabalho. Sendo assim, ao considerar as equações 2.13 e 2.19, a função de sobrevivência e, consequentemente, a função densidade de probabilidade do modelo proposto são definidas, respectivamente por:

$$S_{pop}(y) = (1-\phi) + \phi \exp\left\{ -\exp\left\{ \frac{y - \mathbf{x}^T\boldsymbol{\beta}}{\sigma} \right\} \right\} \tag{3.1}$$

e

$$f_{pop}(y) = \phi \frac{1}{\sigma} \exp\left\{ \left(\frac{y - \mathbf{x}^T\boldsymbol{\beta}}{\sigma} \right) - \exp\left(\frac{y - \mathbf{x}^T\boldsymbol{\beta}}{\sigma} \right) \right\}, \tag{3.2}$$

A estimação dos parâmetros foi feita por meio do método de máxima verossimilhança restrita. Sendo o vetor de parâmetros $\boldsymbol{\theta}$ sob k restrições de inequações lineares, $\mathbf{u}_i^T\boldsymbol{\theta} - c_i \geq 0, i = 1, 2, ..., k$, em que $\mathbf{u}_i^T$ é um vetor $k \times 1$ do tipo $(1, 0, 0, ..., 0)$, assumindo valor 1 na posição em que o parâmetro de interesse se encontra, e c_i são escalares assumindo valores 0 ou 1 de acordo com a restrição de interesse.

A representação do logaritmo da função de verossimilhança sujeito às restrições

lineares é dada por:

$$l_R(\boldsymbol{\theta}, v) = l(\boldsymbol{\theta}) + v \sum_{i=1}^{k} (\mathbf{u}_i^T \boldsymbol{\theta} - c_i), \tag{3.3}$$

em que $v > 0$ é o parâmetro de ajuste e $(\mathbf{u}_i^T \boldsymbol{\theta} - c_i)$ é o conjunto de restrições de inequações lineares.

Esse método de estimação é utilizado considerando o método da função barreira adaptada. Para maiores detalhes consultar Lange (1999) e Fachini (2011).

Neste estudo, as estimativas de máxima verossimilhança com restrição nos parâmetros serão feitas com auxílio do *software R* (R Core Team, 2015) através da função *constrOptim*.

Capítulo 4

Resultados e Discussões

4.1 Análise Descritiva

Para a análise descritiva dos dados, foi feita inicialmente a estimativa de Kaplan-Meier para a função de sobrevivência, como mostra a Figura 4.1, e em seguida a construção do gráfico da curva TTT plot (Figura 4.2) para avaliar o comportamento da função de risco.

A Figura 4.1 mostra indícios de que a função de sobrevivência é imprópria, ou seja, quando $t \to \infty$, a função de sobrevivência não tende a zero, o que indica a existência de uma possível fração de cura entre os dados.

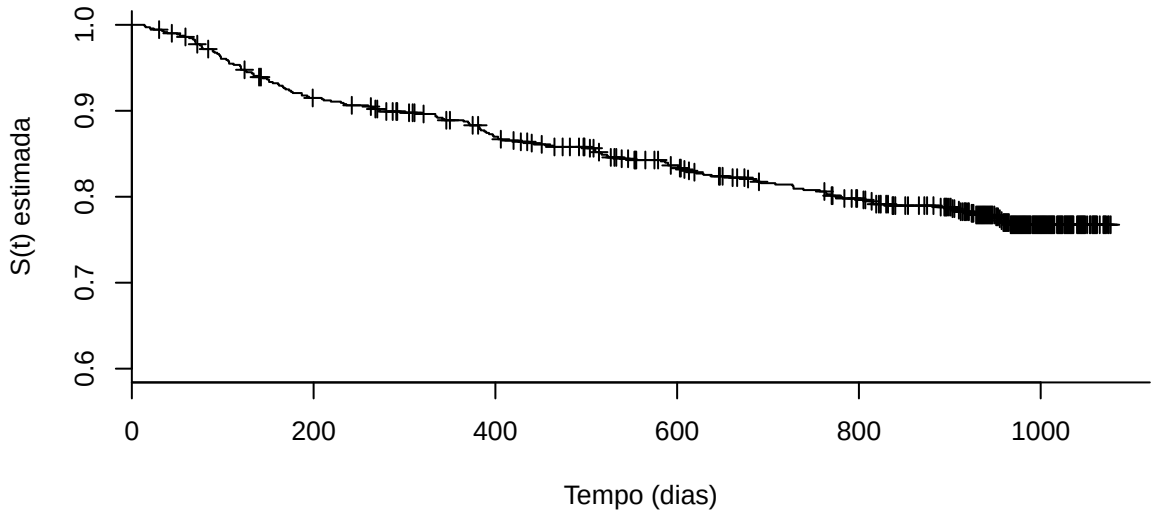

Figura 4.1: Curva estimada pelo método não paramétrico de Kaplan-Meier para os tempos de sobrevivência dos indivíduos com Aids.

A Figura 4.2 mostra que a curva TTT assume uma forma côncava, o que indica que a distribuição que modelará bem os dados será uma função de distribuição de probabilidade

em que a sua função risco assuma a forma estritamente crescente. Por esse motivo, escolheu-se a distribuição Weibull para ser estudada, pois ela permite a forma crescente da função de taxa de falha.

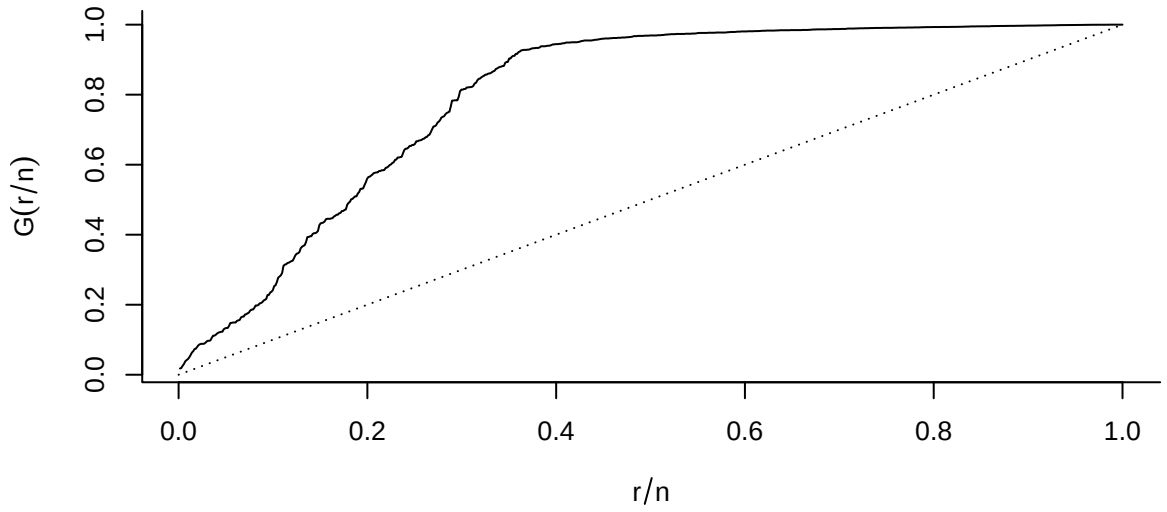

Figura 4.2: Curva do Tempo Total em Teste para os dados dos pacientes com Aids.

Em seguida, foi feita uma análise que leva em consideração a covariável *comprimdia*, presente no banco de dados, que representa o número médio de comprimidos tomados pelos pacientes por dia. A análise de Kaplan-Meier, quando leva em consideração covariáveis, da um indicativo da significância dessas variáveis no modelo de regressão. Assim, para fazer as curvas de sobrevivência não paramétricas estimadas pelo método de Kaplan-Meier para a covariável *comprimdia*, foi feita uma categorização da variável com base na sua mediana, pois essa se apresentava de forma discreta e com poucas observações em alguns valores, então os critérios adotados foram: O primeiro grupo de pacientes foi composto por aqueles que tomam até 5 comprimidos por dia, e o segundo grupo composto por aqueles que tomam mais de 5 comprimidos.

A Figura 4.3, então, mostra as curvas estimadas por esses grupos de pacientes. Ao observa-la, é possível supor que exista diferença entre as curvas, e já que há evidências das curvas não possuírem riscos proporcionais, ou seja, as curvas não são paralelas pois se cruzam em algum momento, também foi feito o teste não paramétrico de Wilcoxon, para verificar se essa covariável tem influência no tempo de sobrevivência dos pacientes.

O teste de Wilcoxon possui as seguintes hipóteses:

$$\begin{cases} H_0\text{: A covariável não influência no tempo de sobrevivência;} \\ H_1\text{: A covariável influência no tempo de sobrevivência.} \end{cases}$$

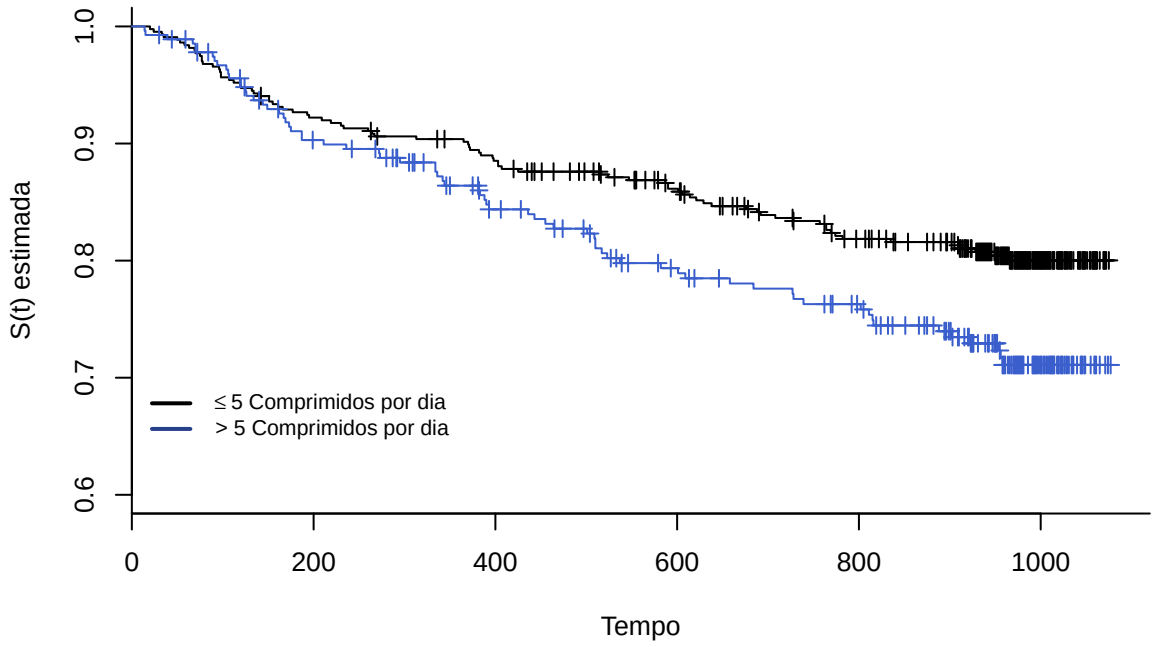

Figura 4.3: Curva estimada pelo método não paramétrico de Kaplan-Meier para os tempos de sobrevivência dos indivíduos com Aids pelo número de comprimidos tomados.

Os resultados obtidos foram:

Tabela 4.1: Resultado do teste de Wilcoxon para a variável *comprimdia*.

Estatística do Teste (χ^2)	p-valor
$5,6$	$0,0176$

Como o p-valor encontrado foi de $0,0176$, a um nível de significância de 5%, faz com que a hipótese nula seja rejeitada, ou seja, existem evidências estatísticas para concluir que o número de comprimidos tomados por dia, influencia no tempo de sobrevivência dos pacientes, ou seja, aparentemente, tomar muitos comprimidos por dia, não aumenta a chance de que não ocorra a falha virológica, como podemos observar novamente na Figura 4.3, em que a curva do Grupo 1, que é a dos pacientes que tomam até 5 comprimidos por dia, tem os tempos de sobrevivência mais elevados.

A seguir, a Figura 4.4 mostra as estimativas das curvas TTT para a covariável *comprimdia*, ao observa-la é fácil perceber que as curvas se assemelham bastante com a da Figura 4.2, que considera todos os tempos. Por isso, ao verificar mais uma vez um comportamento de forma crescente nas curvas da função de risco dos dados, neste trabalho foi considerada a distribuição Weibull em sua modelagem.

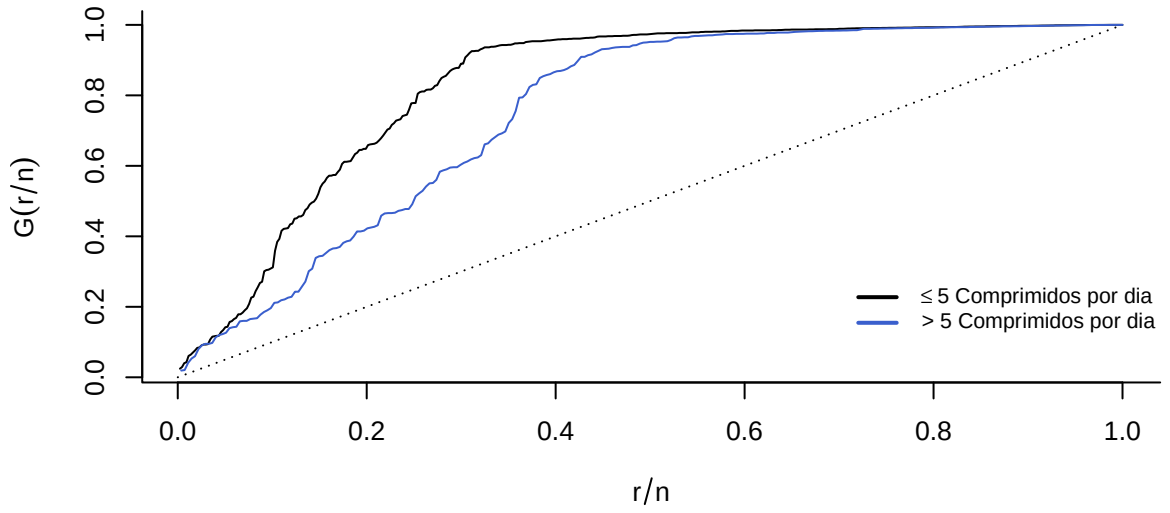

Figura 4.4: Curva do Tempo Total em Teste para os dados dos pacientes com Aids pelo número de comprimidos tomados.

4.2 Modelagem

Após realizar a análise descritiva e ter observado os indícios de fração de cura nos dados, foi feita inicialmente uma comparação da modelagem da distribuição log-Weibull, com e sem fração de cura. A Figura 4.5 apresenta essa comparação das curvas de sobrevivência com base no gráfico de Kaplan-Meier, dessa vez considerando o log dos tempos, já que as funções de sobrevivência também foram feitas dessa forma.

Com isso, pode-se perceber que a Figura 4.5 mostra que a suposição de fração de cura no modelo parece ser evidente, já que a curva com fração de cura parece se encaixar melhor aos dados, e além disso, ao analisar a curva do modelo sem fração de cura observou-se um certo afastamento dos dados no final do estudo, ou seja, aquela certa pré disposição da função de sobrevivência de tender a zero, coisa que, como pode-se perceber, a curva com fração de cura não apresenta.

Em seguida, foram incluídas covariáveis no modelo de regressão proposto através do modelo descrito na seção 3.2, e foram feitas as estimativas de máxima verossimilhança para os parâmetros associados. A Tabela 4.2 apresenta essas estimativas, juntamente com o erro padrão, a estatística Z do teste de significância, descrito na seção 2.4.3, e o p-valor obtido.

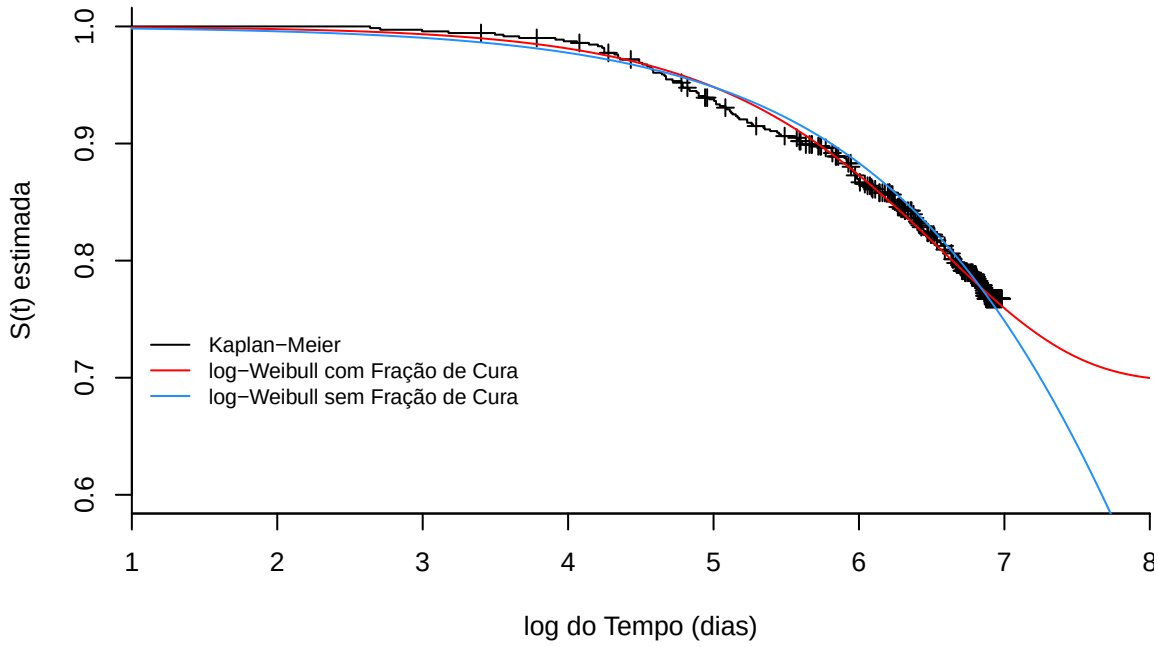

Figura 4.5: Ajuste da curva de Kaplan-Meier e das curvas de sobrevivência da distribuição log-Weibull, com e sem fração de cura, estimadas para os tempos de sobrevivência dos indivíduos com Aids.

Tabela 4.2: Estimativas do modelo de regressão log-Weibull com fração de cura para os dados de pacientes com Aids.

Parâmetros	Estimativas	Erro Padrão	Z	p-valor
σ	$1,0008$	$0,1064$	$-$	$-$
β_0	$7,5677$	$0,6112$	$12,3817$	$< 0,0001$
β_1	$-0,4610$	$0,2319$	$-1,9884$	$0,04677$
ϕ	$0,3733$	$0,0967$	$-$	$-$

Assim, analisando esses resultados, pode-se concluir que a covariável *comprimdia* introduzida no modelo mostrou-se significativa à um nível de significância de 5%, como já era suspeito a partir da Figura 4.3 e do resultado do teste de Wilcoxon.

Agora, para testar se o parâmetro ϕ de cura é significativo no modelo, não é interessante usar o mesmo teste feito anteriormente para os outros parâmetros, já que existe uma restrição no espaço paramétrico de ϕ. Malller e Zhou (1996), propõe então uma pequena modificação do teste da razão de verossimilhanças para testar a significância do parâmetro na fronteira do espaço paramétrico, deve-se assumir que $2(\log(L(\boldsymbol{\theta})) - \log(L(\boldsymbol{\theta}_0))$ tem distribuição Qui-quadrada definida por: $P(X \leqslant x) = \frac{1}{2} + \frac{1}{2}P(\chi^2 \leqslant x)$. O percentil de 95% desta distribuição é $2,71$. Caso o valor da estatística do teste seja menor que $2,71$, não rejeita-se a hipótese nula de que $\phi = 1$. No caso do modelo exposto na Tabela 4.2, tem-se: $2(-548,7358 + 550,2274) = 2,9833 > 2,71$. Portanto, o resultado revela que existem

evidências para rejeitar a hipótese nula de que o parâmetro de cura não é significativo no modelo, confirmando a hipótese já levantada anteriormente apenas com a análise gráfica da Figura 4.5, ou seja, neste estudo há indivíduos não suscetíveis à falha, isto é, indivíduos em que a adesão dos medicamentos se mostrou eficiente, e assim estes não deixarão de fazer efeito, fazendo com que a falha virológica não ocorra.

Contudo, apesar do modelo se encaixar bem aos dados, ao observar a estimativa do parâmetro σ, percebe-se que esta se aproxima de 1, ou seja, seria possível pensar que, na verdade, um modelo com distribuição do Valor Extremo Padrão ou *log*-Weibull Padrão, que considera o parâmetro de escala igual a 1, também seria um bom ajuste para os dados em estudo. Foi verificada então essa hipótese, e os resultados encontram-se na Tabela 4.3.

Tabela 4.3: Estimativas do modelo de regressão log-Weibull Padrão com fração de cura para os dados de pacientes com Aids.

Parâmetros	Estimativas	Erro Padrão	Z	p-valor
β_0	$7,5646$	$0,4589$	$16,4847$	$< 0,0001$
β_1	$-0,4606$	$0,2247$	$-2,0498$	$0,0404$
ϕ	$0,3727$	$0,0671$	$-$	$-$

A Tabela 4.3 mostra que, à um nível de significância de 5%, a covariável *comprimdia* também se mostrou significativa para o modelo considerando $\sigma = 1$. Foi feito novamente o teste da razão de verossimilhanças para verificar a significância do parâmetro ϕ na fronteira do espaço paramétrico para este novo modelo ajustado, então, foi considerada a seguinte estatística do teste: $2(-548,7358 + 552,5195) = 7,5674 > 2,71$. Isso implica que a hipótese nula do teste de que $\phi = 1$ também pode ser rejeitada neste caso, ou seja, o parâmetro de cura também é significativo no modelo com $\sigma = 1$.

Por fim, foi feita uma comparação gráfica dos ajustes desses modelos com a curva de Kaplan-meier utilizando novamente o log dos tempos, para facilitar a comparação entre eles, a Figura 4.6 apresenta essas comparações. Com isso, é possível perceber que os dois modelos são aparentemente bons ajustes quando comparados com seus respectivos modelos sem a inclusão da fração de cura (curvas verde e amarela), porém, sob uma perspectiva mais rigorosa pode-se concluir que o modelo log-Weibul Padrão, ou Valor Extremo Padrão, descrito pela curva azul, parece ser mais adequado para esses dados do que o modelo log-Weibull (curva vermelha), apesar deles estarem claramente bem próximos, contudo a curva vermelha parece superestimar um pouco os tempos ao longo do estudo.

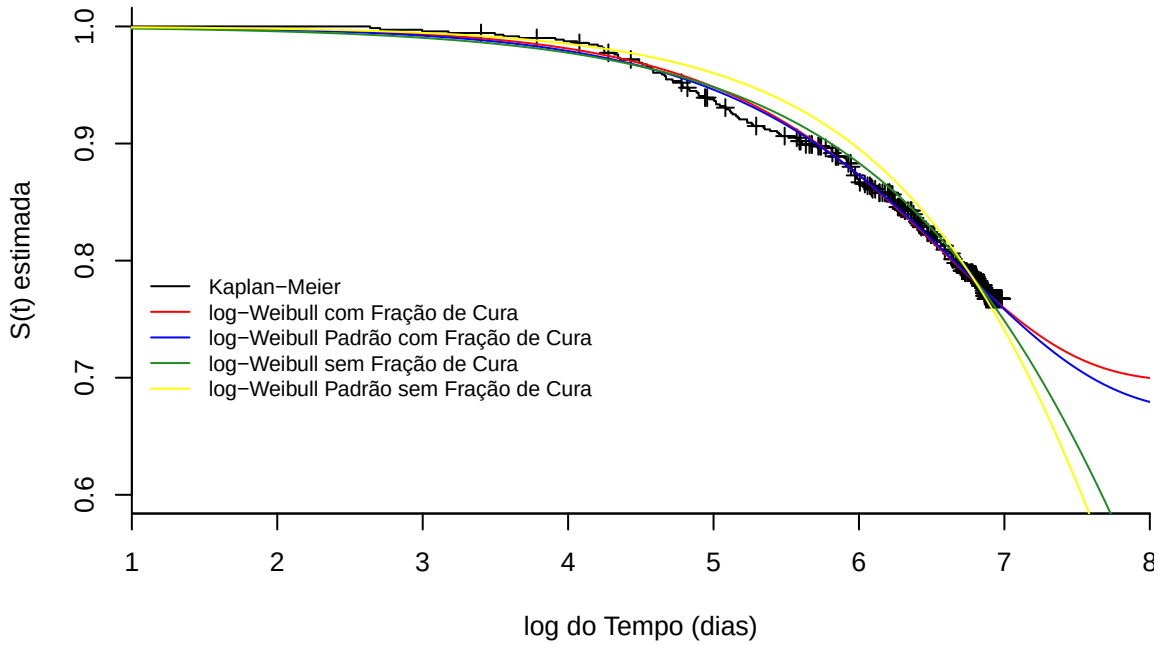

Figura 4.6: Ajuste da curva de Kaplan-Meier e das curvas de sobrevivência da distribuição log-Weibull e log-Weibull Padrão, com e sem fração de cura, estimadas para os tempos de sobrevivência dos indivíduos com Aids.

4.3 Diagnóstico

Para analisar globalmente se os modelos log-Weibull e log-Weibull Padrão, com fração de cura e com a covariável *comprimdia* são bons ajustes para os dados dos pacientes com Aids, foi realizada então uma análise de resíduos considerando os resíduos de Cox-Snell, resíduos *Martingal* e resíduos *Deviance*.

A Figura 4.7 a seguir mostra o gráfico do resíduo de Cox-Snell. A partir dessa análise gráfica, torna-se claro que o modelo log-Weibull com fração de cura é um bom ajuste, já que o primeiro gráfico, que traz a função de sobrevivência dos resíduos vs a função de sobrevivência do modelo exponencial padrão dos resíduos, se aproxima bastante de uma reta, e o segundo gráfico, que traz as mesmas funções, porém representadas de maneira diferente, confirma que o modelo está ajustado.

Em seguida, a Figura 4.8 descreve o comportamento dos resíduos *Martingal* e *Deviance* para o mesmo modelo. Pode-se observar o comportamento aleatório dos dados e que todos os pontos estão bem modelados. Também é possível notar uma tendência a formação de dois grupos, este fato deve-se a categorização da covariável introduzida no modelo.

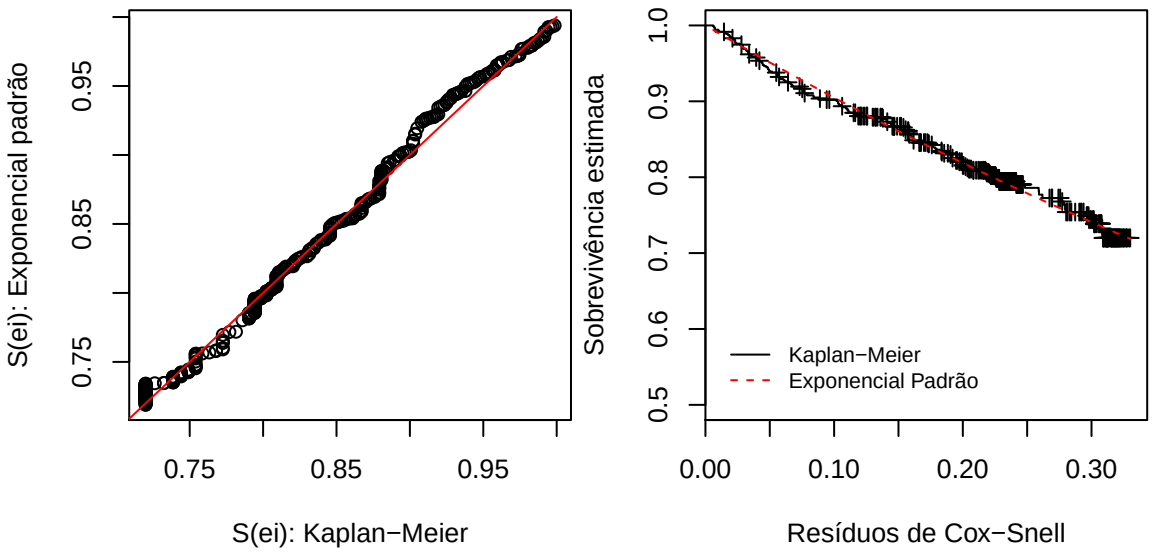

Figura 4.7: Resíduos de Cox-Snell para a distribuição log-Weibull.

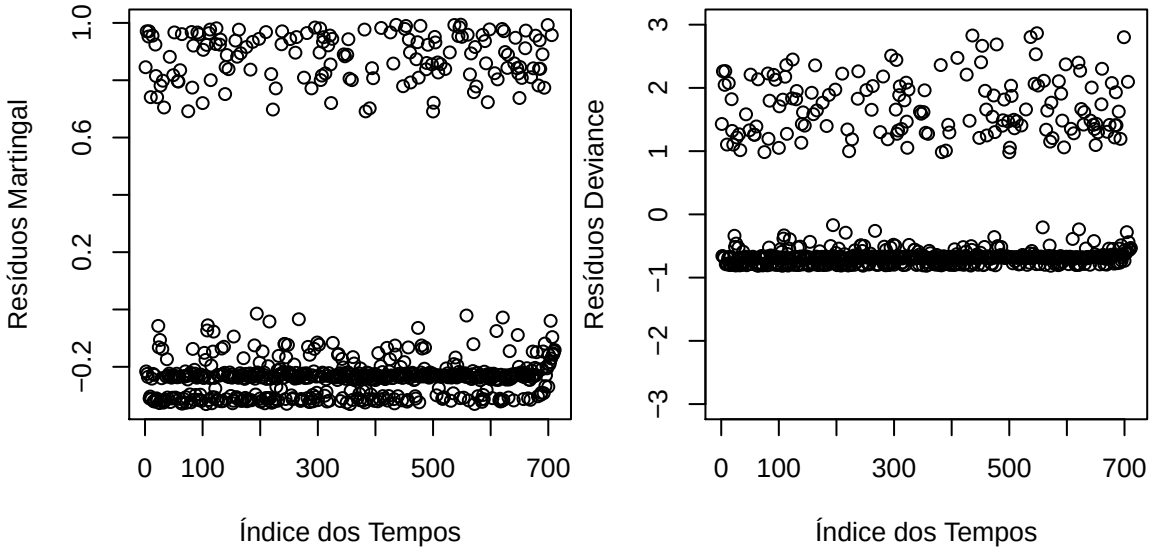

Figura 4.8: Resíduos *Martingal* e *Deviance* para a distribuição log-Weibull.

Agora, a mesma análise foi feita para o modelo log-Weibull Padrão com fração de cura, e os resultados obtidos estão descritos nas Figuras 4.9 e 4.10. Os gráficos obtidos foram tão eficientes quanto os do modelo anterior, pois é fácil perceber que eles são extremamente parecidos, isso significa que tanto o modelo log-Weibull quanto o modelo lo-Weibull Padrão, ambos com fração de cura, são bons ajustes para os dados. A tomada de decisão sobre qual modelo utilizar então, vai de acordo com os critérios do pesquisador, mas, considerando os resultados anteriores e levando em conta a lei da parcimônia, seria possível dizer que a melhor escolha seria ainda o modelo log-Weibull Padrão com fração de cura, por ser o modelo mais simples e mais fácil de trabalhar, a Figura 4.11 mostra então o ajuste desse modelo, juntamente com a estimativa de curados no estudo.

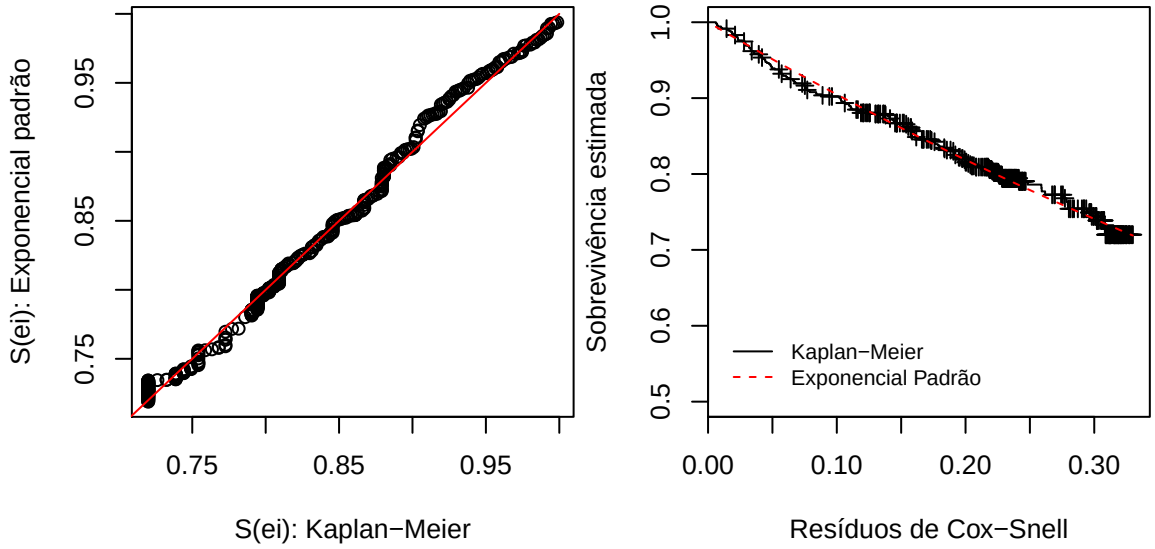

Figura 4.9: Resíduos de Cox-Snell para a distribuição log-Weibull Padrão.

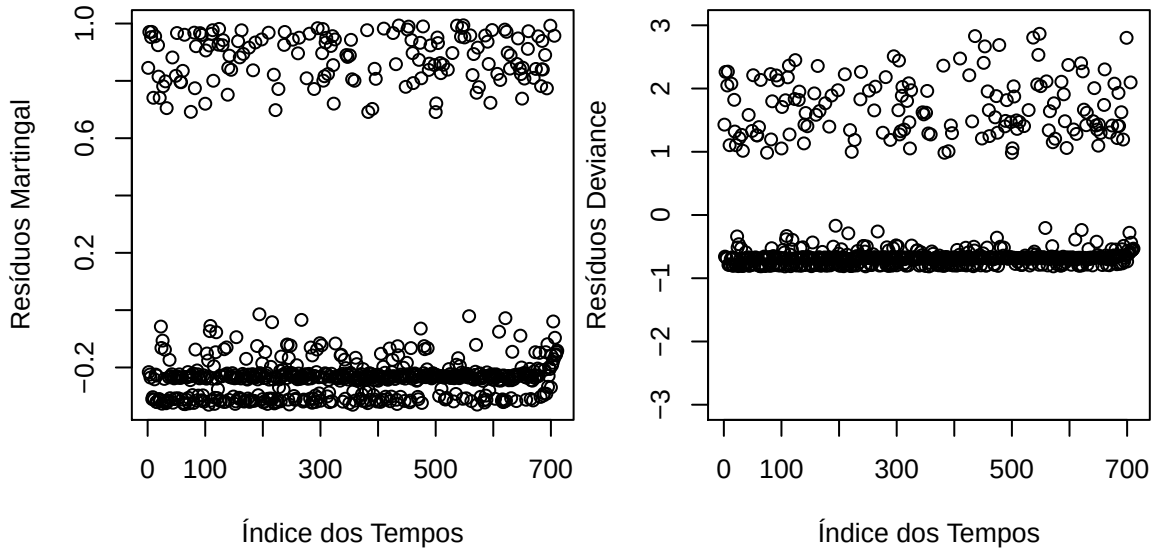

Figura 4.10: Resíduos *Martingal* e *Deviance* para a distribuição log-Weibull Padrão.

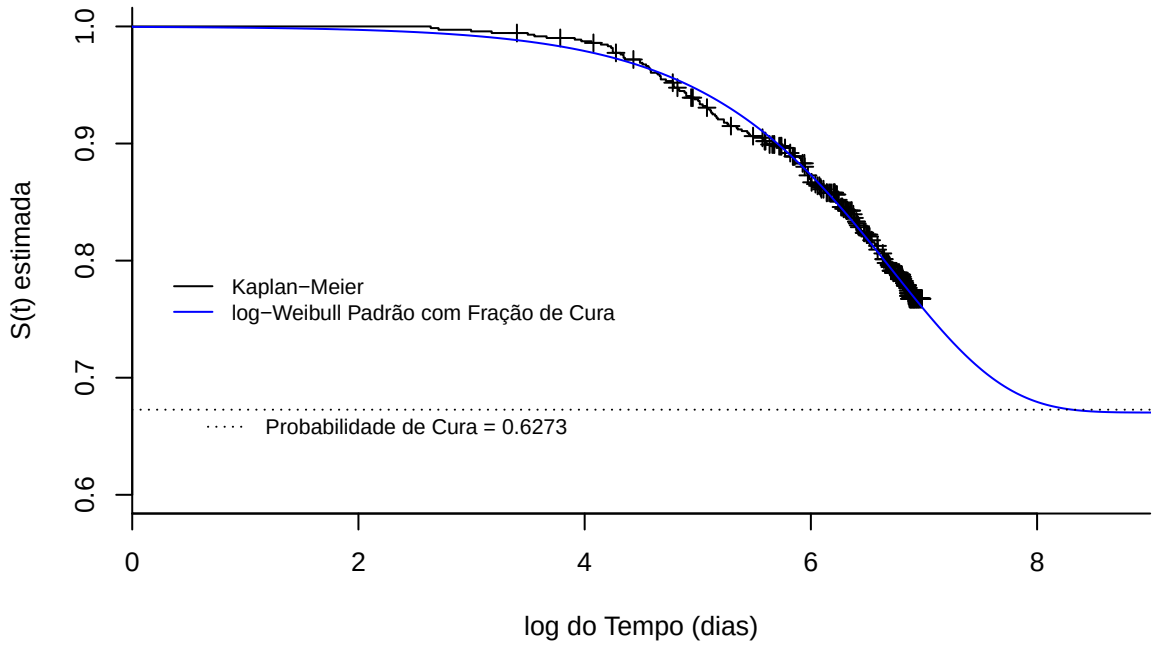

Figura 4.11: Ajuste da curva de Kaplan-Meier e da curva de sobrevivência da distribuição log-Weibull Padrão com fração de cura, estimada para os tempos de sobrevivência dos indivíduos com Aids.

Capítulo 5

Considerações Finais

Neste trabalho foi feito um estudo da distribuição log-Weibull com fração de cura para os dados dos pacientes com Aids que foram submetidos ao tratamento (HAART). Com base nos resultados encontrados foi possível concluir que, como esperado, o modelo proposto, que possui função de risco crescente, é um bom ajuste aos dados.

Mas ao mesmo tempo, ao perceber que o parâmetro de escala do modelo mostrou-se bem próximo de 1, o teste da razão de verossimilhanças entre o modelo log-Weibull com fração de cura e o modelo log-Weibull padrão com fração de cura, e as análises gráficas feitas, mostraram que o modelo log-Weibull Padrão, ou Valor Extremo Padrão, que possui função de risco constante, não somente é também um ótimo ajuste, mas um ajuste mais eficiente.

Além disso, na proposta inicial deste estudo, foi considerado mais um modelo para a análise dos dados. Este modelo, foi construído sob a perspectiva de Farewell (1977) e Maller e Zhou (1996) que introduz covariáveis na estrutura de fração de cura de acordo com a seção 2.4.5, e foram utilizadas as relações descritas nas equações 2.16 e 2.17, contudo, este modelo não se mostrou eficiente devido a uma certa instabilidade computacional, isto é, a convergência na estimação dos parâmetros não foi eficaz, isso levanta uma hipótese de que, para estes dados, a covaríavel não consegue explicar o efeito da fração de cura.

Com base nos resultados do modelo log-Weibull Padrão apresentado na Tabela 4.3 verificou-se que não necessariamente quanto mais comprimidos o paciente tomar por mais tempo ele evitará a falha virológica, muito pelo contrário, o estudo mostra que os maiores tempos de sobrevivência são dos pacientes que tomam uma quantidade menor de comprimidos por dia. Por fim, a estimativa do parâmetro ϕ de cura indica que, dentre as censuras observadas no estudo, é provável que aproximadamente 63% são provenientes de indivíduos curados.

Referências Bibliográficas

Aarset, M. V. (1987). How to identify bathtub hazard rate. *IEEE Transactions on Reliability*, pages 106–108.

Berkson, J. e Gage, R. P. (1952). Survival curve for cancer patients follwing treatment. *Journal of the American Statistical Association*, v. 47:p. 501–515.

Campos, D. P. (2009). *Efeito do Critério de Diagnóstico da AIDS e da Adesão ao Tratamento Anti-Retroviral na Progressão Clínica em HIV/AIDS*. Doutorado, Escola Nacional de Saúde Pública Sergio Arouca, Rio de Janeiro.

Carrasco, J. M. F. (2007). Modelos de regressão log-weibull modificado e a nova distribuição weibull modificada generalizada. Mestrado em estatística e experimentação agronômica, Escola Superior de Agricultura "Luiz de Queiroz", São Paulo, Piracicaba.

Carvalho, M. S., Andreozzi, V. L., Codeço, C. T., Campos, D. P., Barbosa, M. T. S., e Shimamura, S. E. (2005). *Análise de Sobrevivência: teoria e aplicações em saúde*. Fiocruz, Rio de Janeiro.

Carvalho, T. M. (2011). Análise de sobrevivência aplicada ao riscode crédito: ajustes de modelos paramétricos contínuos a dados de tempo discreto. Universidade de Brasília. Instituto de Ciências Exatas. Departamento de Estatística. Trabalho de conclusão de graduação.

Colosimo, E. A. e Giolo, S. R. (2006). *Análise de Sobrevivência Aplicada*. Edgard Blücher, São Paulo. ABE - Projeto Fisher.

Cooner, F., Banerjee, S., Carlin, B. P., e Sinha, D. (2007). Flexible cure rate modeling under latente activation schemes. *Journal of the American Statistical Association*, v. 102:p. 560–572.

de Oliveira, M. L. (2014). Análise de dados de transplante de medula óssea: Proposta do modelo de regressão kumaraswamy-weibull com fração de cura. Universidade de Brasília. Instituto de Ciências Exatas. Departamento de Estatística. Trabalho de conclusão de graduação.

Fachini, J. B. (2011). *Modelos de regressão com e sem fração de cura para dados bivariados em análise de sobrevivência*. Doutorado em estatística e experimentação agronômica, Escola Superior de Agricultura "Luiz de Queiroz", São Paulo, Piracicaba.

Farewell, V. T. (1977). A model for a binary variable with time-censored observations. *Biometrika*, v. 64:p. 43–46.

Garcia, P. N. A. (2013). Aplicação de técnicas de análise de sobrevivência em paciente submetidos à intervenção coronária percutânea. Universidade de Brasília. Instituto de Ciências Exatas. Departamento de Estatística. Trabalho de conclusão de graduação.

Herrmann, L. (2011). Estimação de curvas de sobrevivência para estudos de custo-efetividade. Universidade Federal do Rio Grande do Sul. Instituto de Matemática. Departamento de Estatística. Trabalho de conclusão de graduação.

Ibrahim, J. G., Chen, M.-H., e Sinha, D. (2001). *Bayesian survival analysis*. Springer, New York. 479 p.

Kalbfleisch, J. D. e Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, page 439. Wiley, New Jersey, 2nd edition.

Lange, K. (1999). *Numerical analysis for statisticians*, page 356. Springer, New York.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, page 439. Wiley, New Jersey, 2nd edition.

Leal, C. (2013). Modelo de regressão log-linear com fração de cura. Universidade de Brasília. Instituto de Ciências Exatas. Departamento de Estatística. Trabalho de conclusão de graduação.

Malller, R. e Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. Wiley, New York. 278 p.

Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V. T., e Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, v. 34:p. 541–554.

R Core Team (2015). R: A language and environment for statistical computing. Vienna, Austria. Disponível em: http://www.R-project.org/.

Rizzato, F. B. (2006). Modelos de regressão log-gama generalizado com fraçãp de cura. Mestrado em estatística e experimentação agronômica, Escola Superior de Agricultura "Luiz de Queiroz", São Paulo, Piracicaba.

Santos, T. A. (2013). Modelo de regressão pertencente à família weibull com fração de cura. Universidade de Brasília. Instituto de Ciências Exatas. Departamento de Estatística. Trabalho de conclusão de graduação.

Tsodicov, A. D., Ibrahim, J. G., e Yakovlev, A. Y. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, v. 98:p. 1063–1078.

Anexos

A.1 Distribuição do Valor Extremo

Considere T uma variável aleatória com distribuição Weibull. Deseja-se encontrar a distribuição de $Y = g(T) = \log(T)$, que é dada por:

$$f_Y(y) = f_T(g^{-1}(y))|J|,$$

onde

$$|J| = \frac{dt}{dy}$$

Sabe-se que a função de densidade de probabilidade de uma variável aleatória T com distribuição Weibull é dada por:

$$f_T(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp\left\{ -\left(\frac{t}{\alpha}\right)^\gamma \right\}, t \geq 0$$

Como $Y = \log(T)$,

$$\frac{dy}{dt} = \frac{1}{t} \Rightarrow \frac{dt}{dy} = t$$

mas $t = \exp(y)$

Portanto,

$$\begin{aligned} f_Y(y) &= \frac{\gamma}{\alpha^\gamma} \exp(y)^{\gamma-1} \exp\left\{ -\left(\frac{\exp(y)}{\alpha}\right)^\gamma \right\} \exp(y) \\ &= \frac{\gamma}{\alpha^\gamma} \exp(y)^{\gamma} \exp\left\{ -\left(\frac{\exp(y)}{\alpha}\right)^\gamma \right\} \end{aligned}$$

Fazendo as seguintes substituições: $\gamma = \frac{1}{\sigma}$ e $\alpha = \exp(\mu)$, tem-se que:

$$\begin{aligned} f_Y(y) &= \frac{1}{\sigma \exp(\mu)^{\frac{1}{\sigma}}} \exp(y)^{\frac{1}{\sigma}} \exp\left\{ - \left(\frac{\exp(y)}{\exp \mu} \right)^{\frac{1}{\sigma}} \right\} \\ &= \frac{1}{\sigma} \exp\left(\frac{y-\mu}{\sigma}\right) \exp\left\{ - \exp\left(\frac{y-\mu}{\sigma} \right) \right\} \\ &= \frac{1}{\sigma} \exp\left\{ \left(\frac{y-\mu}{\sigma} \right) - \exp\left(\frac{y-\mu}{\sigma} \right) \right\}. \end{aligned}$$

Logo, a variável aleatória Y tem distribuição do Valor Extremo com parâmetros μ e σ. A representação é dada por $Y \sim ValorExtremo(\mu, \sigma)$.